Trends in the Open Probability

STATCOM

Executive Summary

Questions of Interest

- What are the factors that increase the probability of a subscriber opening a newsletter?
  - What is the best time of day to send a newsletter?
  - Does the email subject affect the open rate?
- How has the COVID pandemic affected the open rate?

Statistical Analysis

- Barplots to visually display trends
- Statistical model to examine how factors interact
- Results from the model to confirm the trends shown in the barplots

Takeaways

- 10:30 am and 5:10 pm seem to be optimal times for sending the newsletter.
- Shorter subject headings (in terms of number of characters) are better.

Trends in the Open Probability

Overview of the Data Used

There are 16,291 unique subscribers in the data set, with 622,614 observations.

We focused on the probability that a given subscriber will open a weekly newsletter within a week of its sent date. If the subscriber opens the newsletter after a week or uses the newsletter to unsubscribe, we consider it a non-open. We examined several factors affecting the open probability:

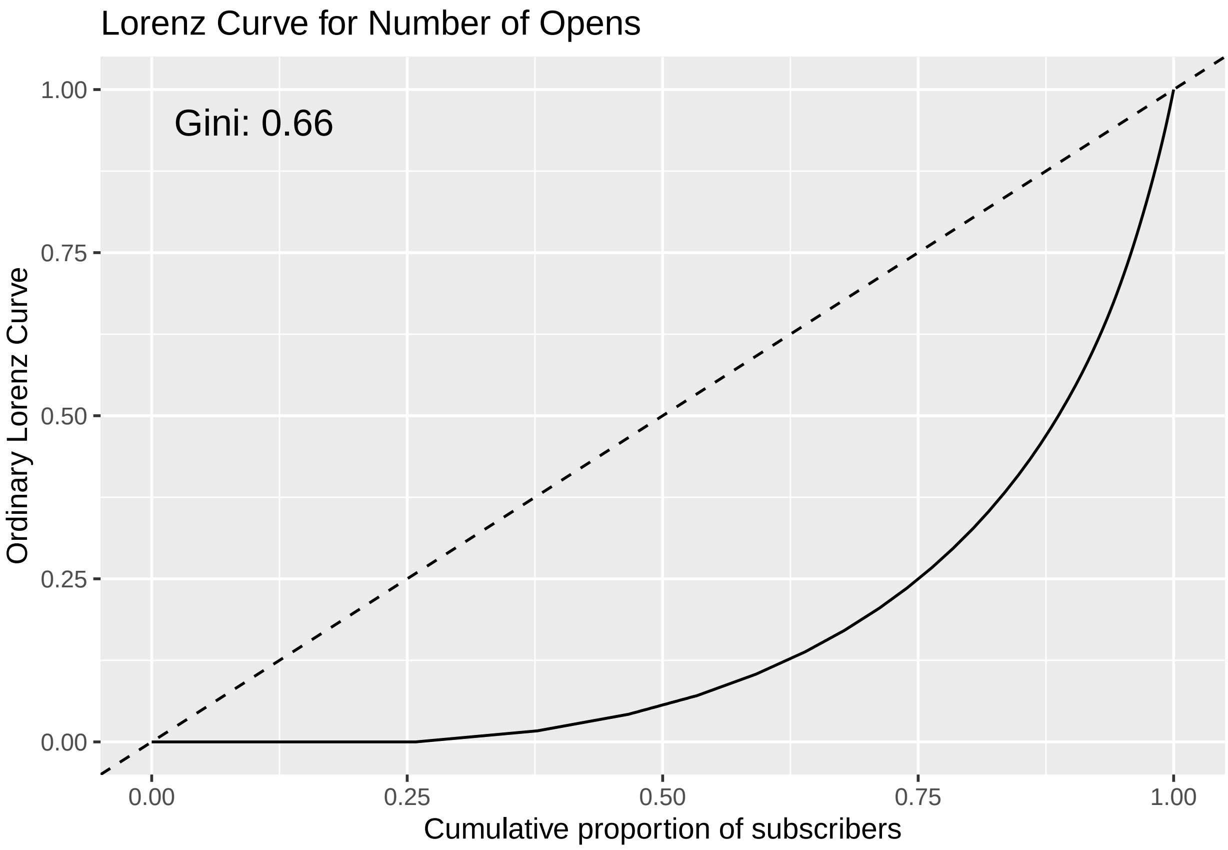
- Trend over the date newsletter was sent out (2019-01-01 to 2020-12-31).
- Whether newsletter was sent before or after start of the COVID pandemic on 2020-03-12.
- Time of day newsletter was sent out (6:30 am to 8:40 pm).
- Length of subject by number of characters.

Below are 10 sample observations from the dataset. The last variable, week_open, is the response variable of interest. 1 indicates that the subscriber opened the newsletter within the week; 0 indicates that the subscriber received the newsletter but didn't open it.

date_sent	subscriberid	covid	mins_since_midnight	subject_length	week_open
2020-07-08 06:51:04	70392049	After	411	65	0
2020-07-21 07:05:50	67928039	After	425	69	0
2020-11-18 12:47:35	71244202	After	767	57	0
2020-12-29 07:30:43	70293552	After	450	39	1
2020-06-17 07:21:09	70811658	After	441	66	0
2020-09-23 07:05:42	71567218	After	425	57	1
2020-07-29 13:30:36	67927884	After	810	65	0
2019-12-03 08:53:03	61625847	Before	533	75	0
2019-12-03 08:53:03	57442167	Before	533	75	0
2020-12-16 07:50:15	71567206	After	470	48	0

Lorenz Curve

The Gini Index ranges from 0 to 1, with 1 being perfect inequality. In this case, the distribution of opens among the subscribers seems unequal; according to the curve, the top 25% of people account for 75% of opens.



Overview of Analysis

For every factor of interest, we plotted a barplot to display any trends.

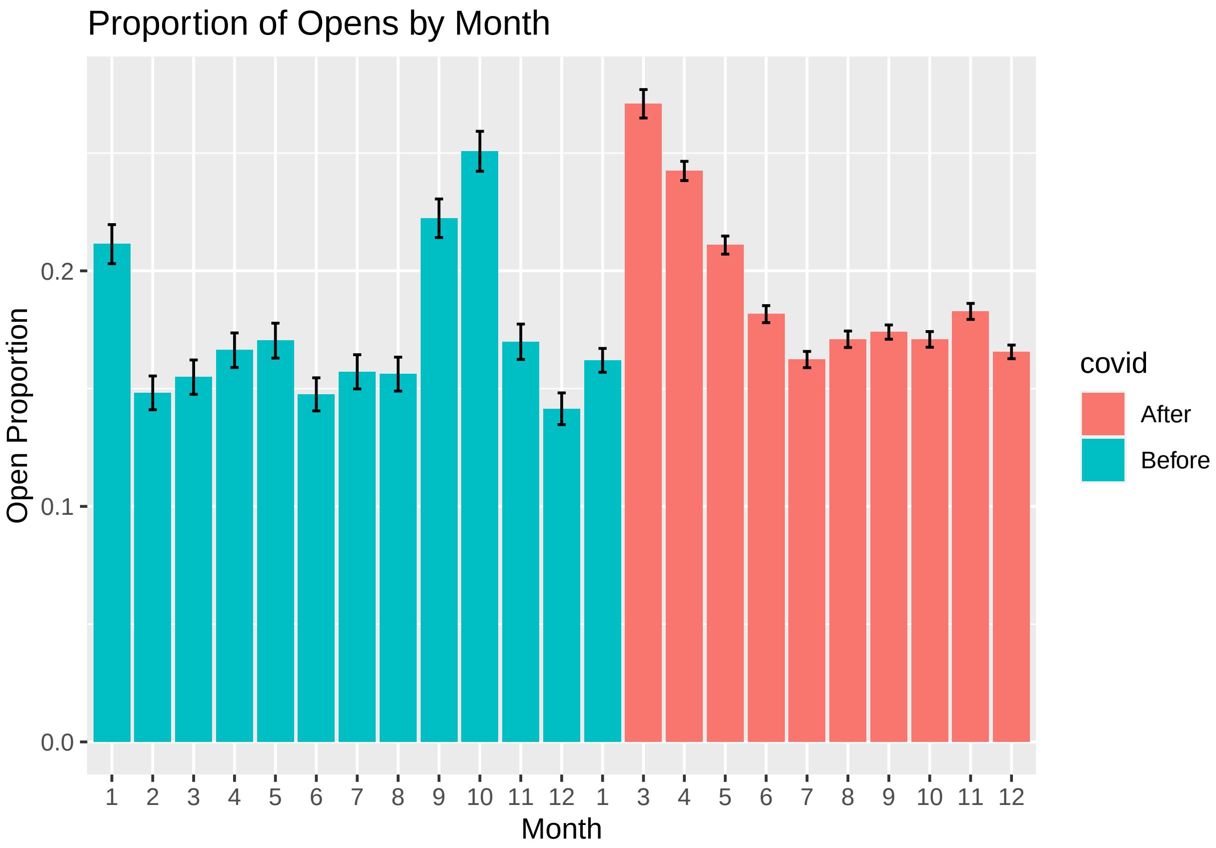
We also fitted a generalized additive mixed model to the data to confirm the trends shown in the barplots. Results from the model are more reliable than just using the barplots, because the model accounts for confounding and dependence between opens for the same subscriber.

The model uses a random sub-sample of 1,629 subscribers (63,897 observations) so it can finish in a reasonable amount of time.

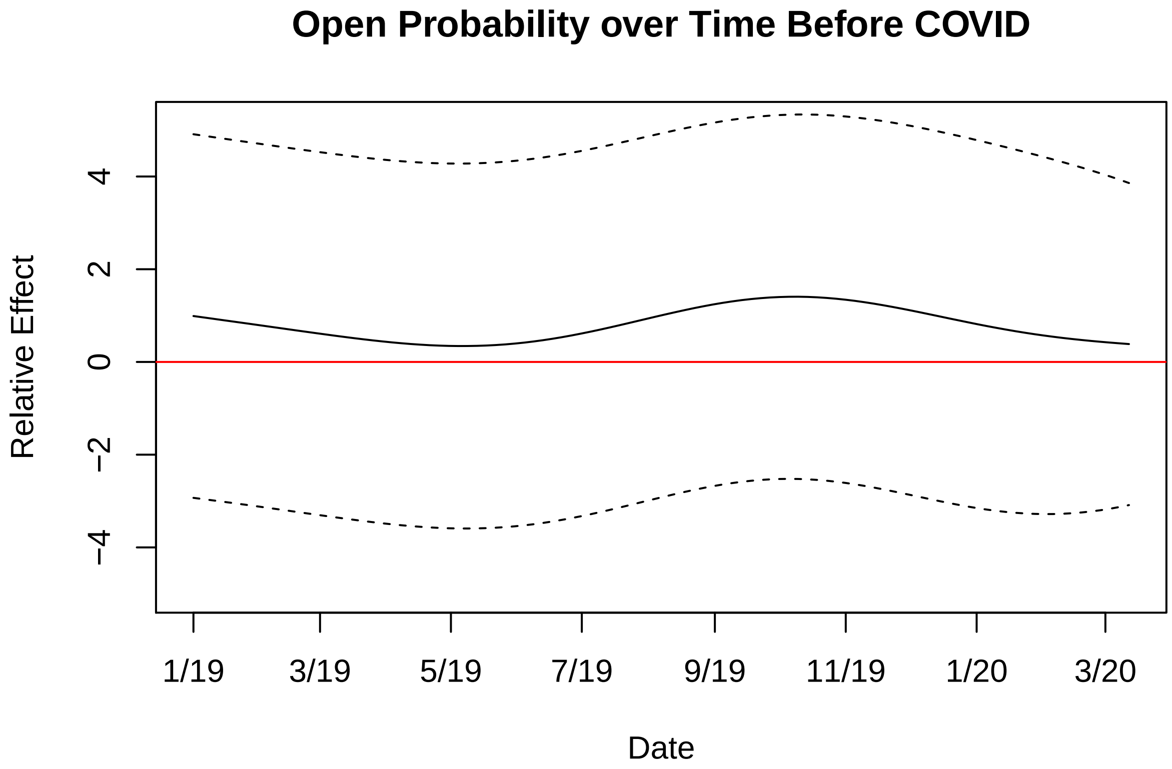
All plots show the confidence intervals of the estimates; two standard errors above and below the estimates are indicated on top of the bars in the barplots and by the dashed lines or shading in the line graphs. The true value can be expected to lie within two standard errors from the estimate.

Trend over Date

The below barplot shows the proportion of subscribers that opened the newsletter, given the month the newsletter was sent to them.

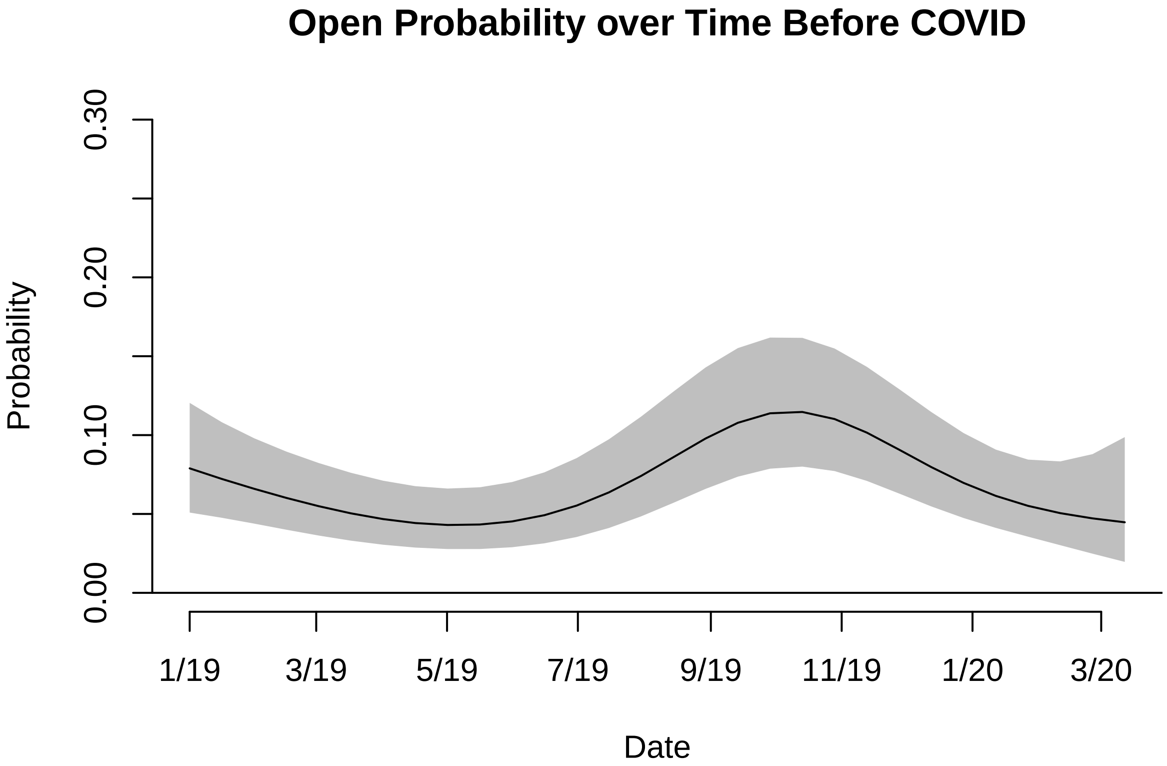


The following plot shows the relative effect of the date the newsletter is sent out on the open probability (a negative relative effect corresponds to a decrease in probability, and a positive relative effect corresponds to an increase in probability) before the pandemic. The red line at zero indicates no effect. There appears to be a seasonal trend.

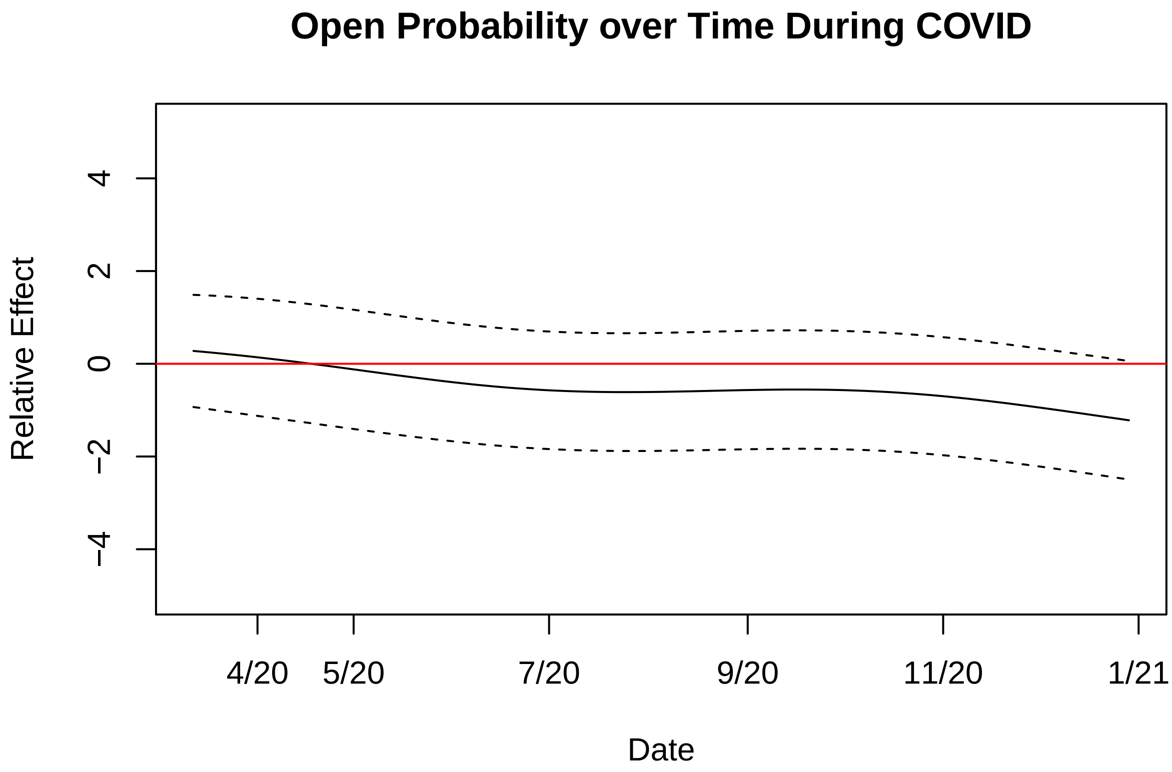


The above plot shows the partial effect of the date alone, without considering other covariates. The following plot shows the actual estimated probabilities over time under the following specific scenario:

- The newsletter was sent out at 10:30 am.
- The newsletter has the median subject length of 66 characters.

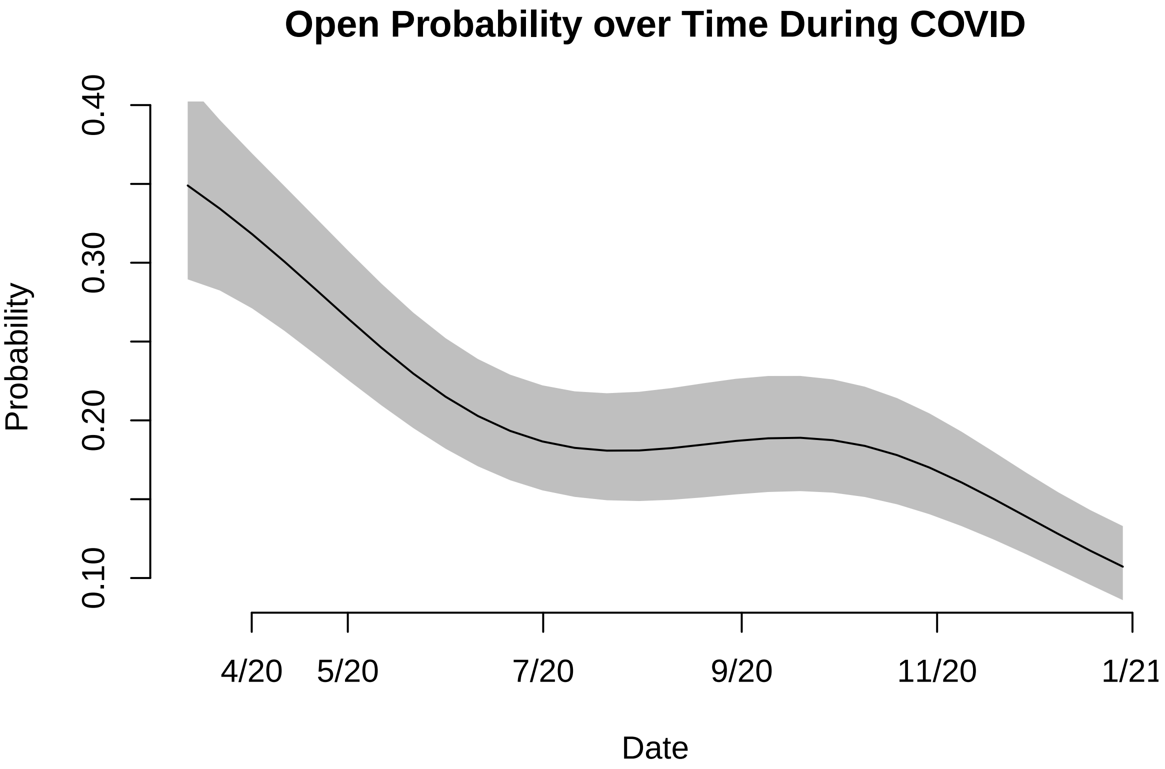


After the pandemic, there is a downward trend in open probability.

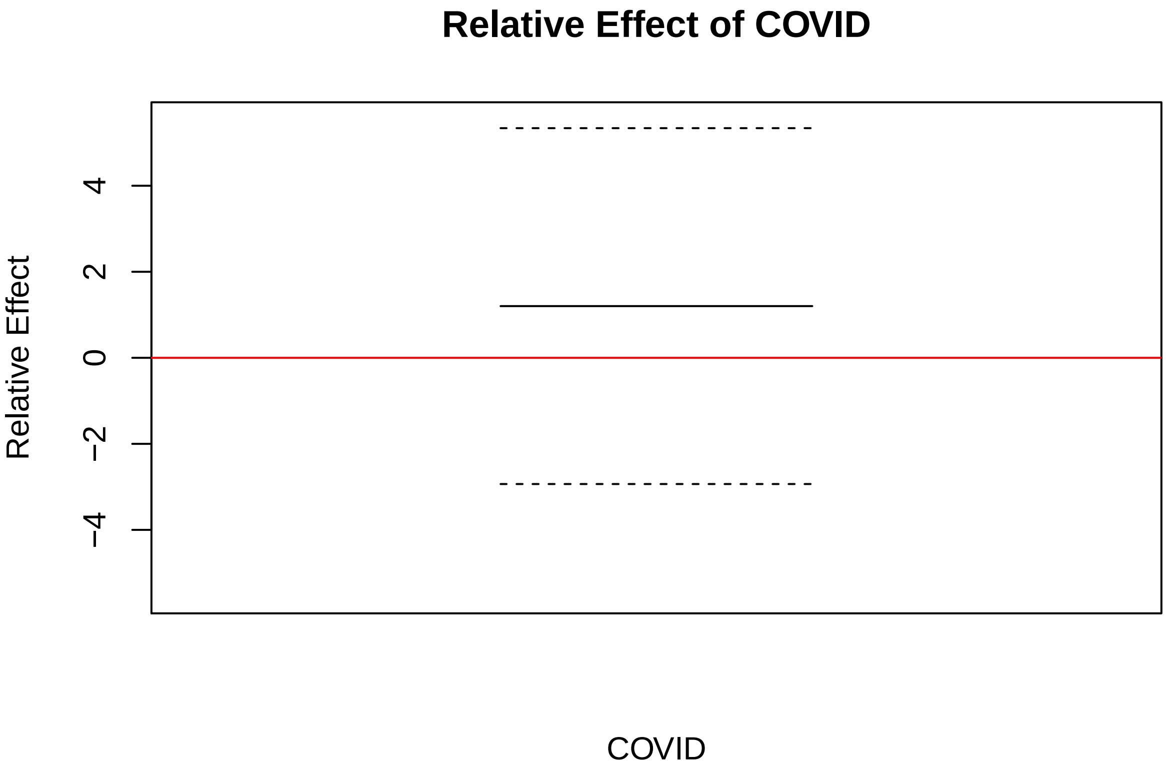


The above plot shows the partial effect of the date alone, without considering other covariates. The following plot shows the actual estimated probabilities over time under the following specific scenario:

- The newsletter was sent out at 10:30 am.
- The newsletter has the median subject length of 66 characters.

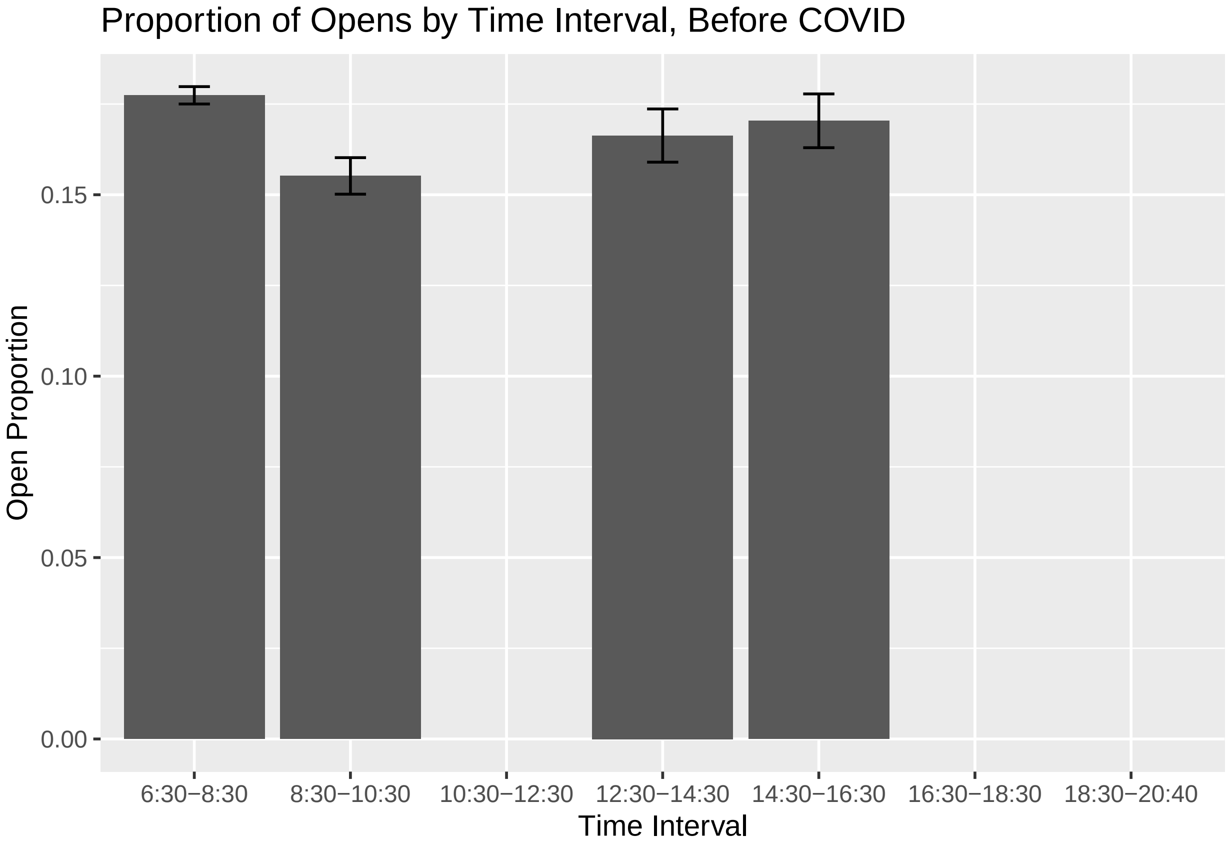


Below shows the relative effect of COVID, i.e. whether the newsletter was sent after the pandemic started. It appears that the open probability rises after the pandemic starts, but the effect is not significantly different than before the pandemic (see the dashed standard error bars). However, COVID significantly affects how the open probability varies by date or hour of day the newsletter was sent.

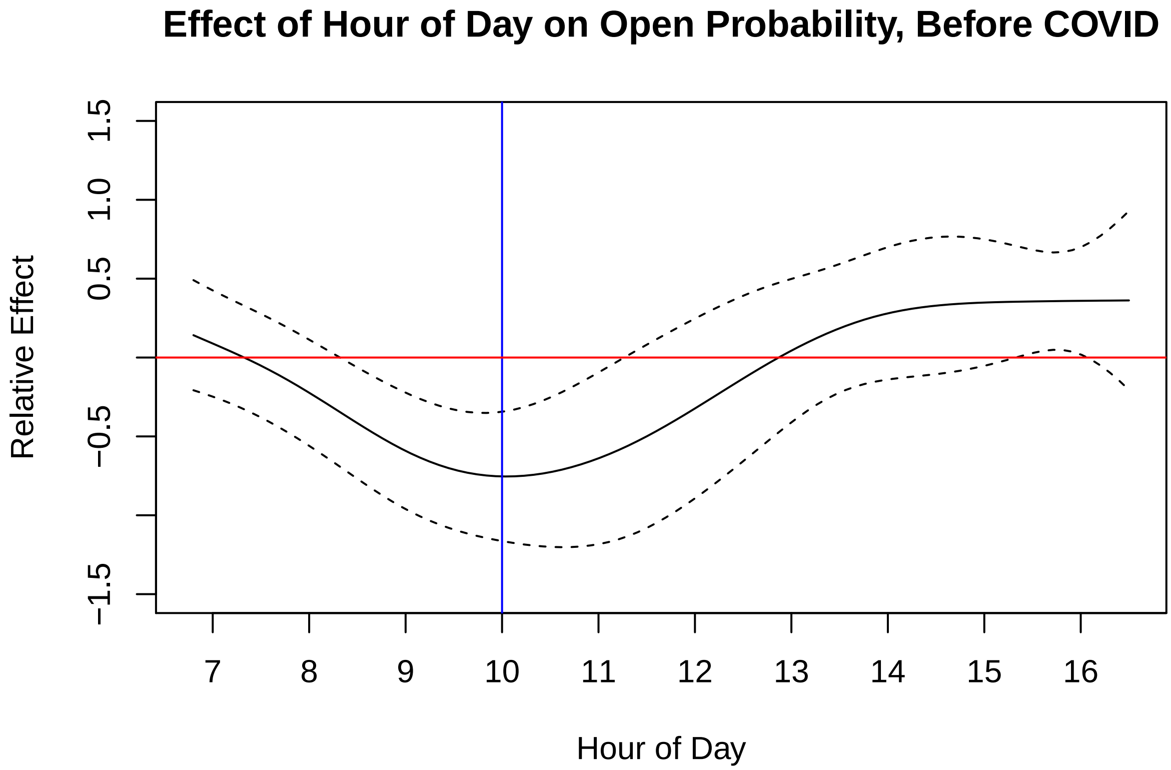


Time of Day Trend, Before COVID

The below barplot shows the proportion of subscribers that opened the newsletter sent before the pandemic, given that the newsletter was sent to them within a specific time interval. There were no newsletters sent in three of the time intervals, so the bars are absent.

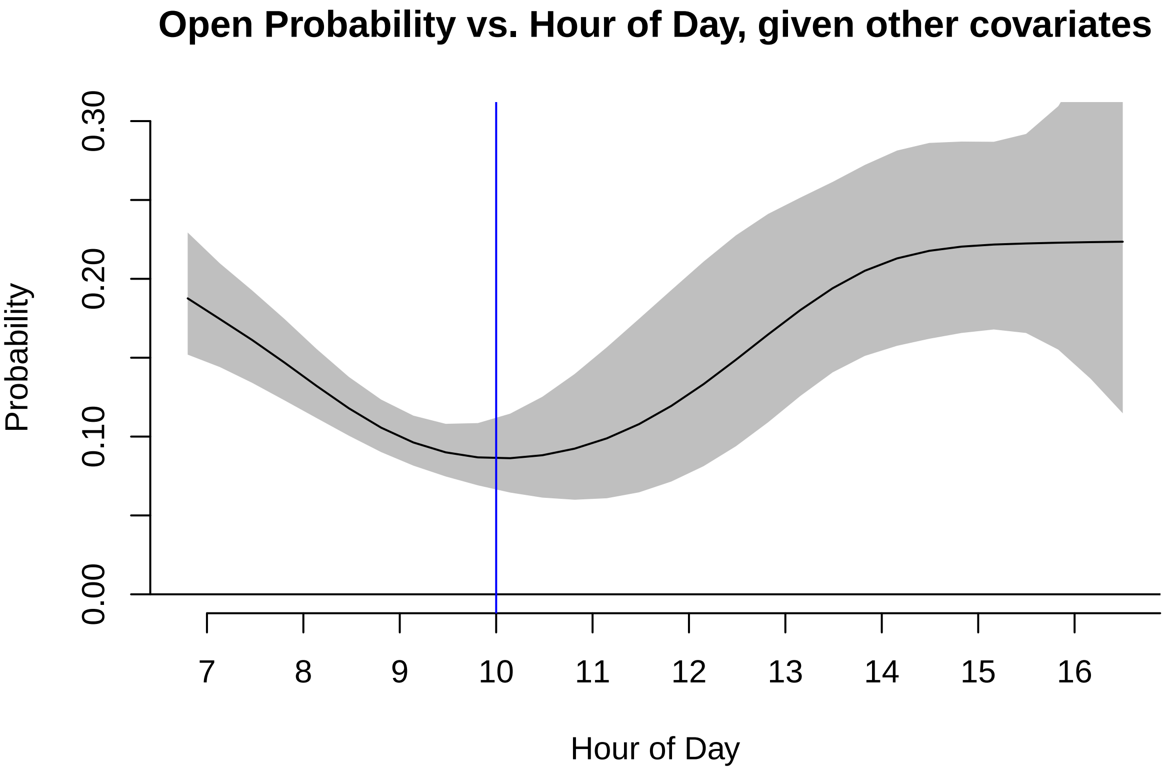


The following plot shows the relative effect of the time of day the newsletter is sent out on the open probability (a negative relative effect corresponds to a decrease in probability, and a positive relative effect corresponds to an increase in probability) before the pandemic. It appears that there is a dip in the open probability at about 10 am.



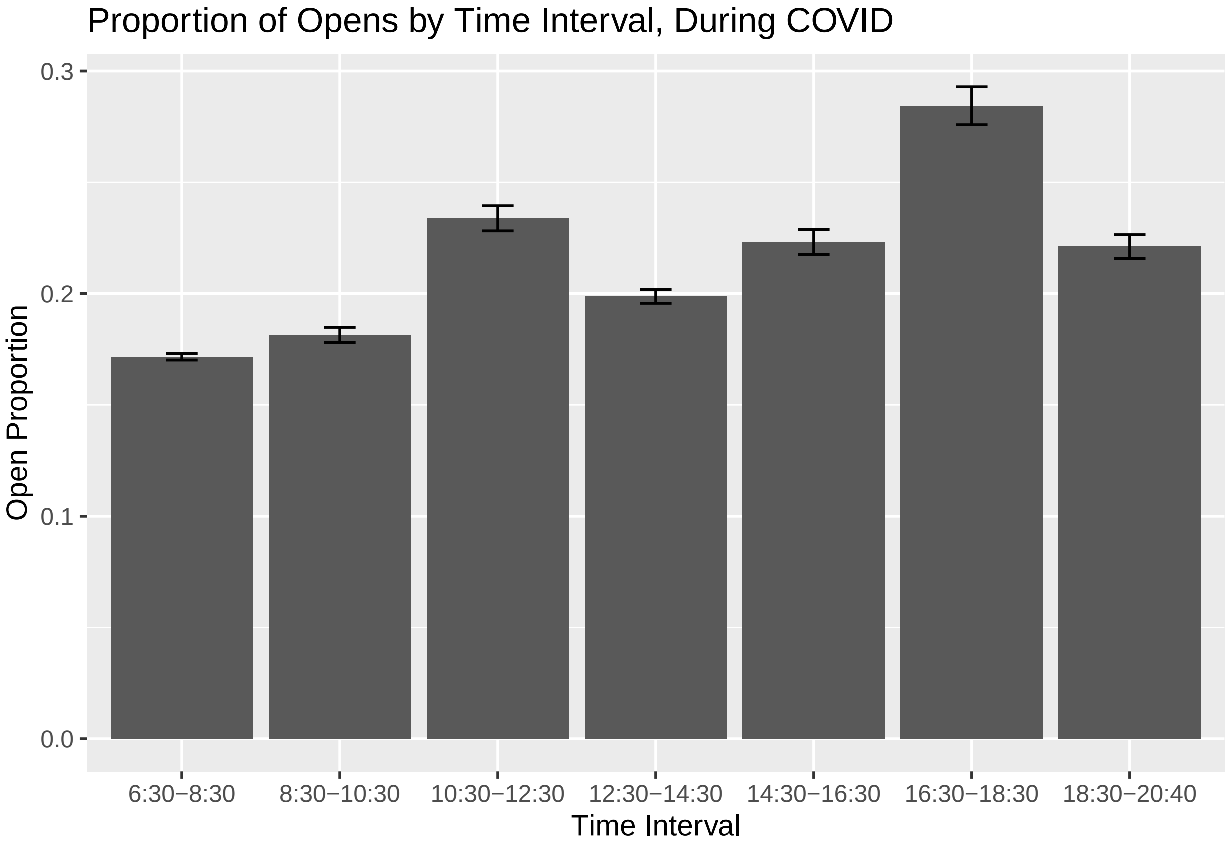
The above plot shows the partial effect of the time of day alone, without considering other covariates. The following plot shows the actual estimated probabilities by time of day under the following specific scenario:

- The newsletter was sent out on December 1, 2019.
- The newsletter has the median subject length of 66 characters.

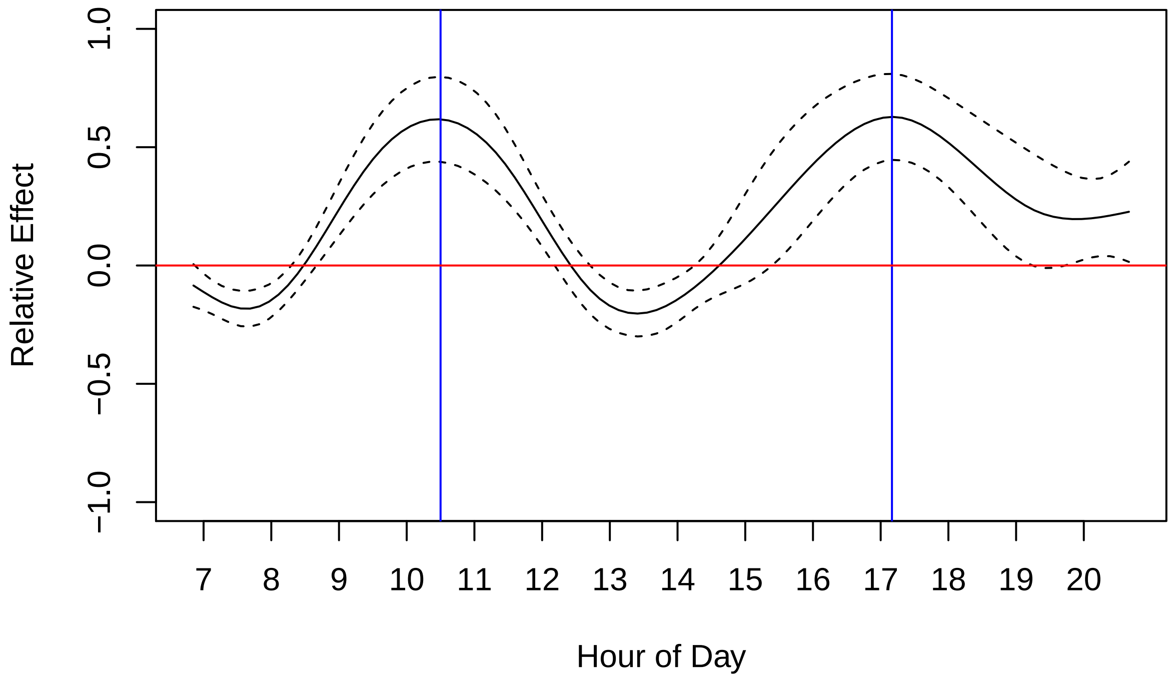


Time of Day Trend, During COVID

The below barplot shows the proportion of subscribers that opened the newsletter during the pandemic, given that the newsletter was sent to them within a specific time interval. The barplot suggests that there are two time intervals with higher open proportions.



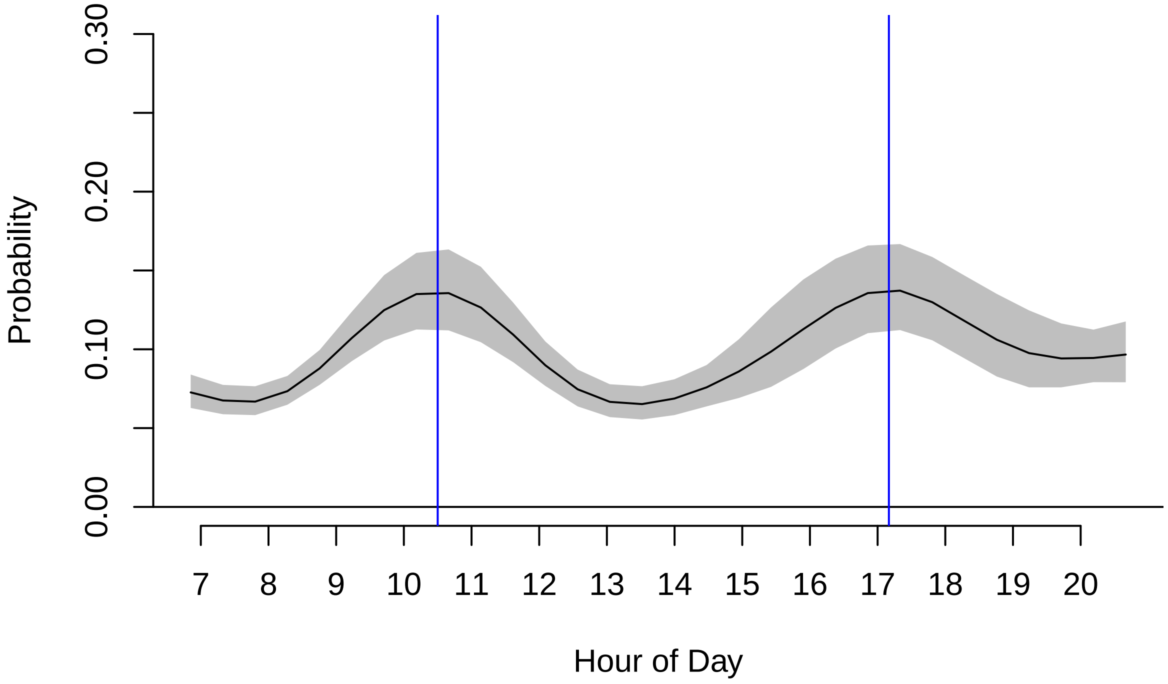
The following plot shows the relative effect of the time of day the newsletter is sent out on the open probability during the pandemic. It appears that the optimal times are about 10:30 in the morning and 17:10 in the evening.



Effect of Hour of Day on Open Probability, During COVID

The above plot shows the partial effect of the time of day alone, without considering other covariates. The following plot shows the actual estimated probabilities by time of day under the following specific scenario:

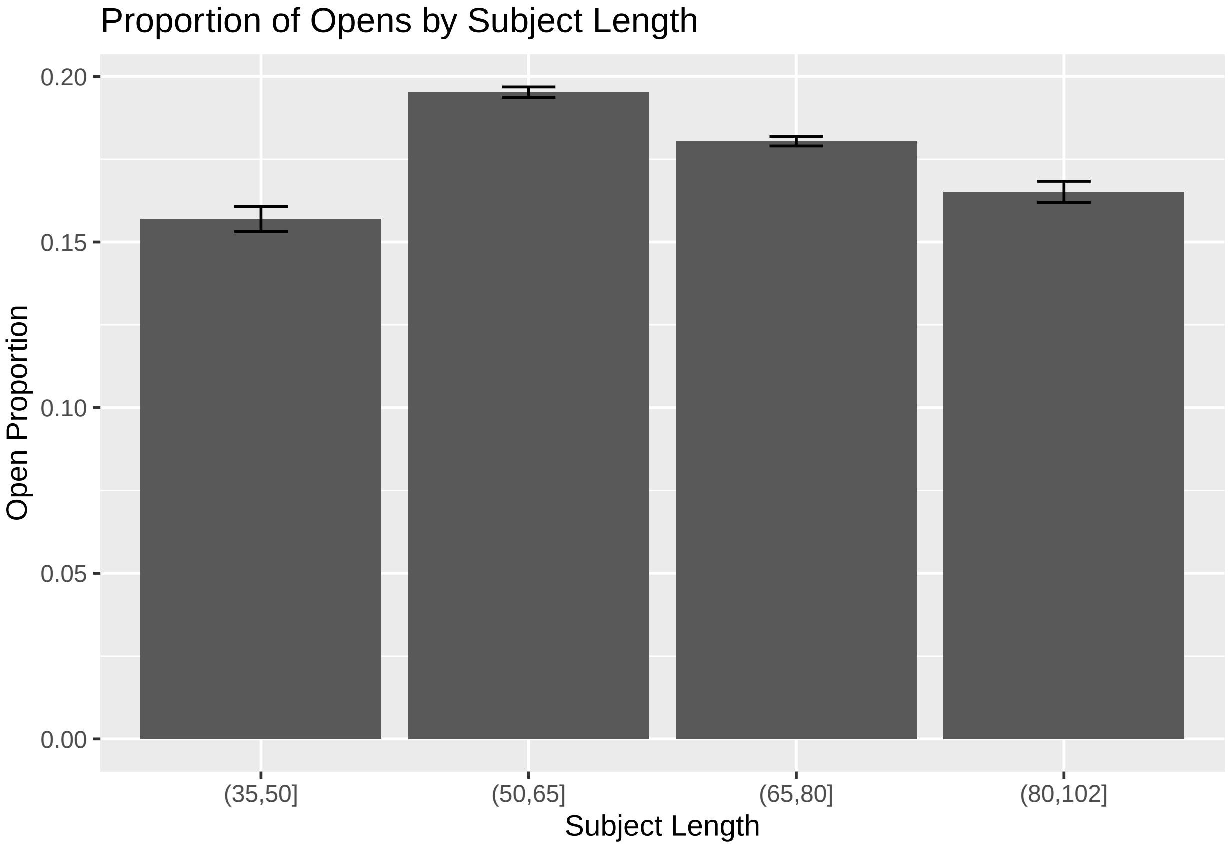
- The newsletter was sent out on December 1, 2020.
- The newsletter has the median subject length of 66 characters.



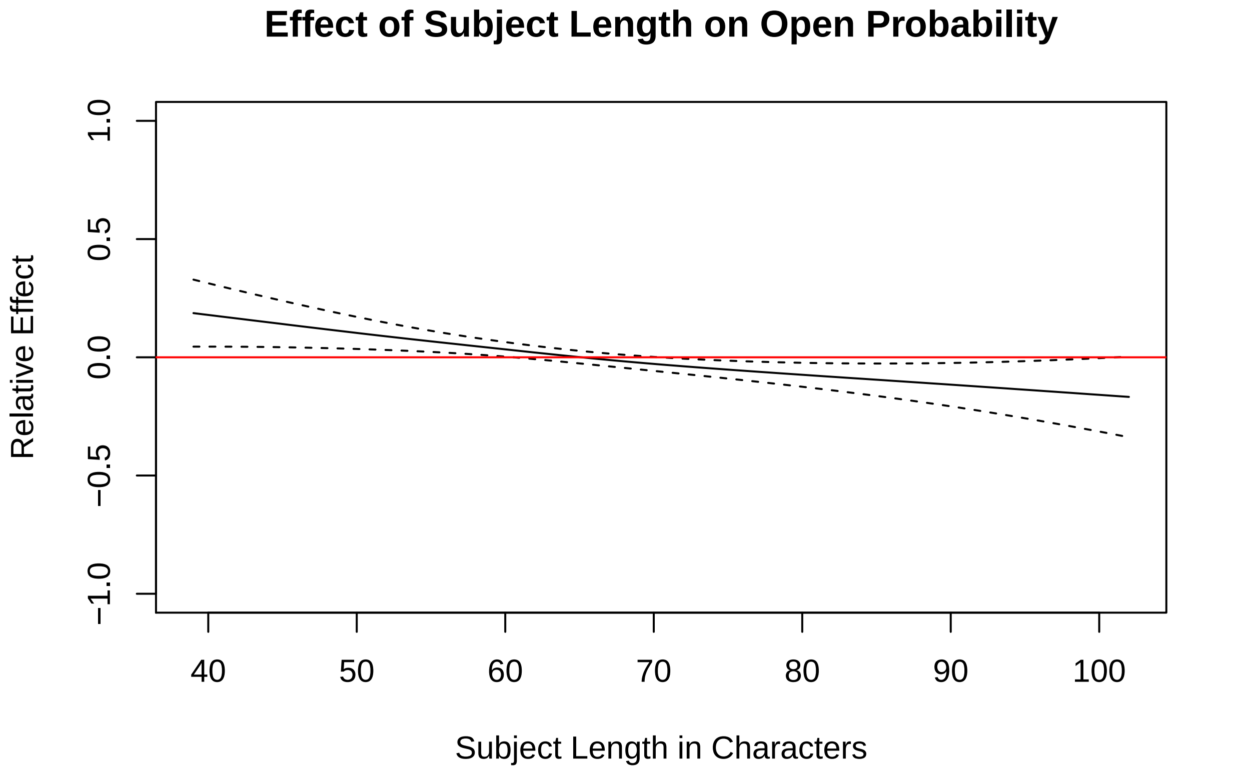
Open Probability vs. Hour of Day, given other covariates

Subject Length Trend

The below barplot shows the proportion of subscribers that opened the newsletter, given that the subject length was within a specific interval.

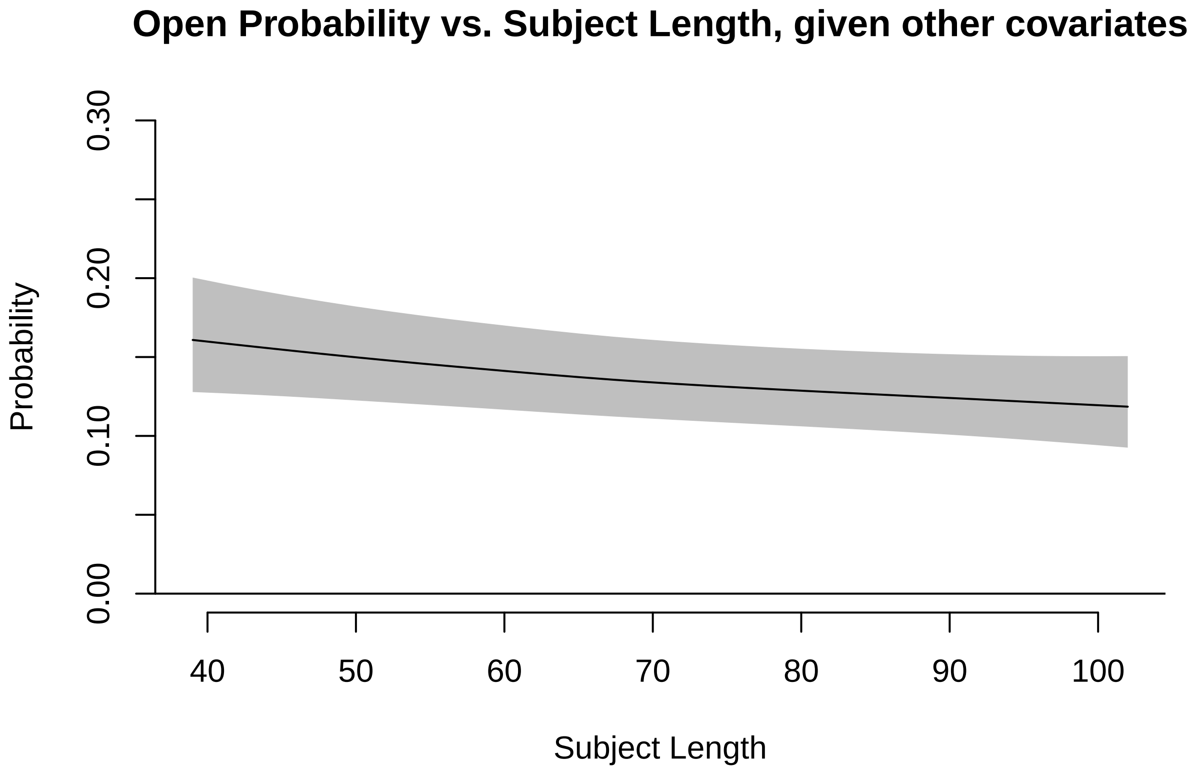


The following plot shows the relative effect of the subject length on the open probability. There appears to be a downward trend in open probability as the subject length increases.



The above plot shows the partial effect of the subject length alone, without considering other covariates. The following plot shows the actual estimated probabilities by subject length under the following specific scenario:

- The newsletter was sent out on December 1, 2020.
- The newsletter was sent out at 10:30 am.


Open Probability vs. Subject Length, given other covariates
Probability
Subject Length